Perceptual separation of simultaneous vowels: Within and across-formant grouping by F_0

John F. Culling and C. J. Darwin
Laboratory of Experimental Psychology, University of Sussex, Falmer, Brighton, East Sussex BN1 9QG, United Kingdom

(Received 30 June 1992; revised 29 January 1993; accepted 22 February 1993)

Six experiments explored why the identification of the two members of a pair of diotic, simultaneous, steady-state vowels improves with a difference in fundamental frequency (ΔF_0). Experiment 1 confirmed earlier reports that a ΔF_0 improves identification of 200-ms but not 50-ms duration "double vowels"; identification improves up to 1 semitone ΔF_0 and then asymptotes. In such stimuli, all the formants of a given vowel are excited by the same F_0, providing listeners with a potential grouping cue. Subsequent experiments asked whether the improvement in identification with ΔF_0 for the longer vowels was due to listeners using the consistent F_0 within each vowel of a pair to group formants appropriately. Individual vowels were synthesized with a different F_0 in the region of the first formant peak from in the region of the higher formant peaks. Such vowels were then paired so that the first formant of one vowel bore the same F_0 as the higher formants of the other vowel. These across-formant inconsistencies in F_0 did not substantially reduce the previous improvement in identification rates with increasing ΔF_0's of up to 4 semitones (experiment 2). The subjects' improvement with increasing ΔF_0 in the inconsistent condition was not produced by identifying vowels on the basis of information in the first-formant or higher-formant regions alone, since stimuli which contained either of these regions in isolation were difficult for subjects to identify. In addition, the inconsistent condition did produce poorer identification for larger ΔF_0's (experiment 3). The improvement in identification with ΔF_0 found for the inconsistent stimuli persisted when the ΔF_0 between vowel pairs was confined to the first formant region (experiment 4) but not when it was confined to the higher formants (experiment 6). The results replicate at different overall presentation levels (experiment 5). The experiments show that at small ΔF_0's only the first-formant region contributes to improvements in identification accuracy, whereas with larger ΔF_0's the higher formant region may also contribute. This difference may be related to other results that demonstrate the superiority of resolved rather than unresolved harmonics in coding pitch.

PACS numbers: 43.71.Es, 43.66.Hg

INTRODUCTION

Listeners can recognize more easily the speech of one speaker when it has a different fundamental frequency (F_0) from other speech occurring at the same time. Evidence has come from two types of experiments. First, intelligibility of LPC-resynthesized sentences masked by similar connected discourse is improved by a difference in F_0 between the two voices (Brokx and Nooteboom, 1982). Second, pairs of simultaneous steady-state vowels ("double-vowels") are identified more accurately when they have different rather than the same F_0's (Scheffers, 1983; Zwicker, 1984; Chalikia and Bregman, 1989; Assmann and Summerfield, 1990).

A number of computational models have been designed to exploit differences in F_0 between voices in various ways. These models have been tested both on continuous speech (Parsons, 1976; Weintraub, 1985; Stubbs and Summerfield, 1990) and on double vowels (Scheffers, 1983; Assmann and Summerfield, 1990; Meddis and Hewitt, 1992). Although the different processes employed by the programs have modeled the available experimental data with varying degrees of success, little experimental evidence has been collected that illuminates the actual processing employed by the human auditory system.

It is possible to distinguish two different ways in which differences in fundamental frequency (ΔF_0's) could help separate competing voices and facilitate the identification of speech sounds when F_0 differs across speakers: (1) segregation of harmonics of different F_0's within the same frequency region, and (2) grouping of distinct spectral regions, such as those around successive formant peaks, which are excited by a common F_0 (Broadbent and Ladefoged, 1957). Each of the computational models above identifies the F_0's which are present and then attempts to group harmonics from any part of the spectrum which share each of those F_0's. In the case of Scheffers' (1983) model and Assmann and Summerfield (1990)'s "place" model, the resolution of the initial frequency analysis is sufficiently poor at high frequencies that the F_0's of the vowels are only likely to be distinguished in the region of the first formant peak. These models can, therefore, only group harmonics of common F_0 in that region. The other, more successful models are more likely to distinguish the F_0's underlying higher formant peaks and so may also be able to perform across-formant grouping.

 0001-4966/93/063454-14$06.00

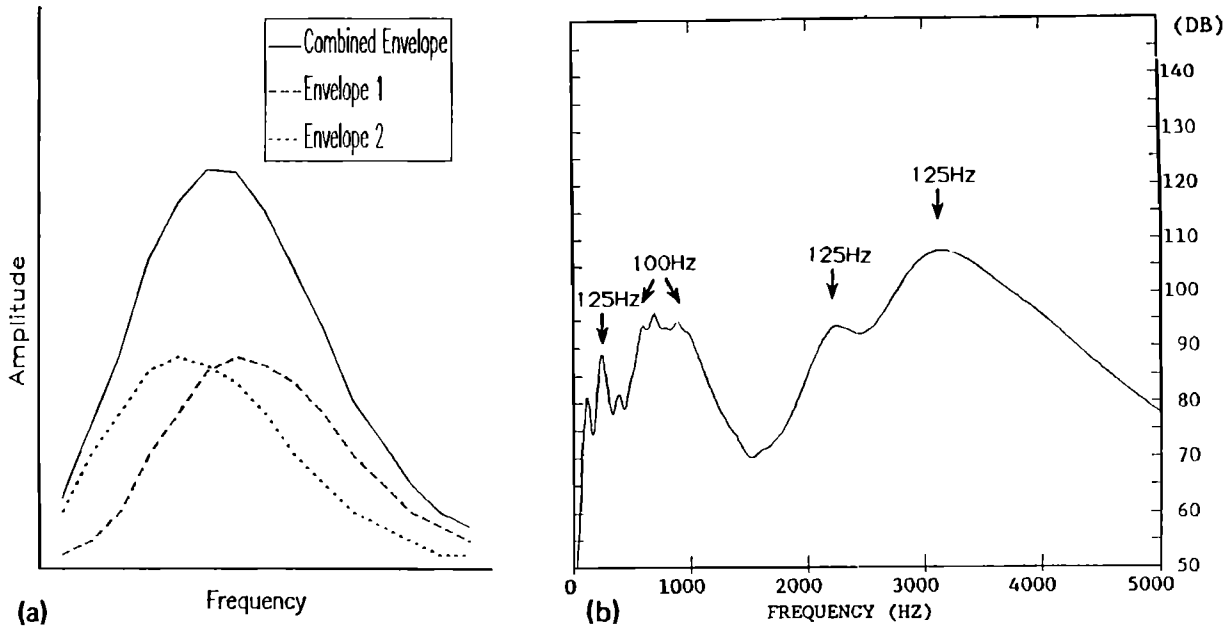

FIG. 1. Two mechanisms of vowel separation: (a) schematic illustration of two overlapping formants merging in the combined spectrum to form a single peak; (b) cochlear excitation pattern of the vowel pair /i/+/ɑ/ with the F_0's of each formant marked.

Grouping of harmonics of each F_0 within the same frequency region may improve formant-frequency estimation. Where formants from competing vowels share the same region of the spectrum, the selection of two different series of harmonics may help reveal the frequencies of the two formant peaks. Figure 1 (left panel) shows schematically two formant-like shapes which sum to make a shape with a single peak. For vowels on the same F_0 only this single peak would be available for phonetic interpretation, but for vowels on different F_0's the two constituent peaks could be derivable from the spectral envelopes of the segregated harmonic series.

Where formants do not overlap, or where overlapping formants have already been separated by component grouping, the F_0's of the harmonics which excite them might then be used to group them with other formants which share the same F_0 in "across-formant grouping" (Broadbent and Ladefoged, 1957). Figure 1 (right panel) shows a cochlear excitation pattern, derived using filters with ERB bandwidths (Moore and Glasberg, 1983) for the pair of vowels /i/+/ɑ/. At low frequencies the pattern resolves a number of individual frequency components, but, given that the auditory system can interpolate between these components to estimate the formant envelopes, the pattern shows five clear formant peaks. These peaks could be grouped in various ways to form different percepts, but the F_0's marked show that the two formants just below 1 kHz come from a different source (the /ɑ/) from the other formants. If the auditory system is able to label and segregate formants which are excited by different F_0's then this mechanism would clearly be valuable in perceptually separating competing voices.

Studies which have used ΔF_0's between formants, in order to explore the role of F_0 in combining formants with common F_0 and in segregating formants excited by different F_0's, have had difficulty in demonstrating any tendency for listeners to group/segregate formants according to their F_0's (Cutting, 1976; Darwin, 1981). Cutting found that listeners had no difficulty in identifying synthetic speech syllables which contained two formants, synthesized on F_0's which differed by as much as 10.2 semitones. Darwin (1981) found similarly robust phonetic labeling for vowels composed of three formants, each with different F_0's (expt. 1). In order to ensure that this result did not come about through subjects identifying vowels from individual formants, Darwin went on to use formant trajectories which formed different diphthongs in different combinations (expt. 2), thus forcing subjects to combine information from the two formants in order to derive a particular diphthong percept. He found no evidence that different F_0 glides (140 Hz rising to 180 Hz and 120 Hz falling to 80 Hz) on two formants of these diphthongs could disrupt diphthong identification. Darwin then looked into the question of whether F_0 can nonetheless affect the grouping of formants into competing perceptual organizations (expt. 3). He prepared four diphthong formants; each of two first formants gave a unique diphthong percept when presented in combination with each of two second formants. He presented all four formants simultaneously in such a way that two pairs of formants could be grouped either by ear of presentation or by common F_0 glide, but found no conclusive evidence in listeners' reported diphthong percepts for grouping of either kind. Finally, however, Darwin (1981, expt. 4), and subsequently, Gardner *et al.* (1989) found that the 2nd formant of a four formant synthetic syllable (/ru/) was excluded from subjects' phonetic percept to produce a different perceived syllable (/li/) when it had been synthesized on a different F_0 from the rest of the formants. Even in this case however the effect occurred only at larger ΔF_0's ($\cong$4.4 semitones)

than those necessary for detection of a second source (<1 semitone) and much larger than those which show higher scores in Scheffers' double-vowel paradigm ($\frac{1}{4}$ and $\frac{1}{2}$ semitone).

One possible explanation of the relative ineffectiveness of ΔF_0's in segregating formants is that the human listener has a tendency to recombine potentially separated sounds which together produce a phonetically meaningful unit. If so, in the cases of Cutting (1976) and Darwin (1981, expt. 1), in which each formant of a speech sound was excited by a different F_0, the formants in isolation were not identifiable speech sounds and only in combination did they acquire phonetic significance. Listeners consequently heard the combined sound. In the case of the "ru/li" paradigm (Darwin, 1981, expt. 4; Gardner *et al.*, 1989), in which the second of four formants is excited by a different F_0, the second formant must be perceptually excluded for the remaining formants to be perceived as /li/, leaving an isolated and meaningless second formant, heard as a buzzing sound. Listeners consequently heard /li/ only when there was a large ΔF_0.

This phonetic constraint explanation may, however, be inconsistent with the results of Darwin (1981, expt. 3), in which two first formant ($F1$) and two second formant ($F2$) sounds were employed. Each $F1/F2$ combination formed a different diphthong, but synthesizing a given $F1/F2$ pair with one F_0 glide and the other pair with the other F_0 glide had no effect upon the perceived combination of sounds when all four formants were presented simultaneously. Here there was competition between two organizations, each of which groups all the components into phonetically meaningful units, yet evidence for across-formant grouping was not found.

The experiments of Cutting, Darwin, and Gardner *et al.* have addressed the role of across-formant grouping as a cue and found only weak effects requiring large ΔF_0's. The second potential mechanism, improved formant frequency estimation, may therefore be responsible for the large improvements in double-vowel identification that occur with the introduction of small ΔF_0's. The present experiments set out to investigate the mismatch between the data from double-vowel experiments and from formant-grouping experiments by using across-formant inconsistencies in F_0 in a double-vowel paradigm.

To do this, vowels were synthesized with a discrete change in F_0 between $F1$ and $F2$. These vowels were then combined in such a way that the first formant of each vowel had the same F_0 as the higher formants of the competing vowel. If across-formant grouping affects double-vowel identification then an inconsistency in F_0 across the first two formants (which are the most influential in vowel identification (Petersen and Barney, 1952), should confuse the listener and produce a decrement in performance.

Before tackling the mechanisms of perceptual separation by ΔF_0, experiment 1 replicated two basic double-vowel phenomena, as exemplified by the results of Assmann and Summerfield (1990).

(1) For vowels of 200-ms duration, there is an improvement in accuracy of identification with increasing ΔF_0 which asymptotes above 1 semitone ΔF_0.

(2) For vowels of 50-ms duration, there is no improvement in identification accuracy with increasing ΔF_0, but a dip in accuracy is found at $\frac{1}{4}$ semitone ΔF_0.

TABLE I. Formant frequencies and 3-dB bandwidths used to synthesize the five individual vowels.

	Vowel					
Formant	/i/	/ɑ/	/u/	/ɜ/	/ɔ/	Bandwidth
$F1$	250	650	250	450	350	90
$F2$	2250	950	850	1250	750	110
$F3$	3050	2950	1950	2650	2850	170
$F4$	3300	3300	3300	3300	3300	250
$F5$	3850	3850	3850	3850	3850	300

I. METHODS COMMON TO EACH EXPERIMENT

A. Stimuli

Following Assmann and Summerfield (1990), the experiments used the five British–English tense vowels, /i/, /ɑ/, /u/, /ɜ/, and /ɔ/. The vowels were synthesized on a VAX 11/780 computer using a program which added together sine waves with amplitudes and phases determined by the transfer function produced by the cascade configuration of Klatt (1980)'s parallel/cascade speech synthesizer. The formant frequencies and bandwidths were identical to those of Assmann and Summerfield and are given in Table I, but the relative intensity of the /ɑ/ vowel was 6 dB lower than in their experiment to reduce the range of intensities among the constituent vowels. The vowels were synthesized with 10-ms raised cosine onset and offset ramps and on various F_0's, the lowest of which was 100 Hz.

Double vowels were made by digitally adding waveforms. In each double-vowel stimulus, one vowel always had an F_0 of 100 Hz, the other had the same or a higher F_0. Since no vowel was ever paired with itself (even on a different F_0), there were ten phonetically different vowel pairs, requiring different responses from the subject. Each pair was represented by two stimuli at each ΔF_0, each with a different allocation of F_0's to vowels (making 20 vowel combinations at each ΔF_0).

B. Procedure

Subjects were first given the individual vowels to identify. Each vowel was played once in a random order with no feedback and subjects who made more than two errors were required to repeat the practice session until they achieved this criterion.

Sounds were played at a 10-kHz sampling rate via a 12-bit digital-to-analog converter through an antialiasing filter (4.5-kHz low pass) and presented to subjects over Sennheiser HD414 headphones in a sound-attenuating booth. The presentation levels of constituent vowels at 100 Hz F_0 lay in the range 77–85 dB(A). Subjects were instructed to register their two responses sequentially by pressing one of five keys, marked "EE," "AR," "OO,"

"ER," and "OR." The subjects were able to press the same key twice, but were aware that the stimuli always contained two different vowels. No feedback was given.

C. Data analysis

Following Scheffers (1982,1983), Zwicker (1984), and Assmann and Summerfield (1990), we calculated for each stimulus type the percentage of trials on which both vowels were correctly identified.

II. EXPERIMENT 1

A. Stimuli

Each of the five vowels was synthesized with 50- and 200-ms overall durations on six different F_0's: 100, 101.46, 102.93, 105.95, 112.46, and 125.99 Hz (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 equal-tempered semitones, respectively, above 100 Hz). The 100-Hz version of each the five vowels was combined with each of the other four vowels at each of the six F_0's to give 120 different double-vowel stimuli at each duration (20 vowel combinations $\times 6$ ΔF_0's).

B. Procedure

Thirteen subjects, with no declared hearing problems and inexperienced in double-vowel experiments, attended an hour-long session. After identifying the single vowels in the practice, they heard two experimental blocks of stimuli, one of 50-ms stimuli and one of 200-ms stimuli. Six of the subjects listened first to the 50-ms stimuli and seven to the 200-ms stimuli. Within each block, subjects heard each of the 120 stimuli twice in a randomized order.

C. Results

Figure 2 shows that the identification of the 200-ms stimuli improves markedly as ΔF_0 increases from zero to $\frac{1}{4}$ and $\frac{1}{2}$ semitone ΔF_0's, and then asymptotes. By contrast, identification of the 50-ms stimuli does not improve with increasing ΔF_0, in fact there is a slight drop in performance at $\frac{1}{4}$ and $\frac{1}{2}$ semitones.

The pattern of data is broadly similar to that of Assmann and Summerfield (1990); both durations yield similar accuracy of identification for zero ΔF_0, but identification rates improve asymptotically with increasing ΔF_0 in the 200-ms condition and show no improvement in the 50-ms condition. Compared with Assmann and Summerfield's data, improvement in the 200-ms condition is greater for $\frac{1}{4}$ semitone ΔF_0 and asymptotes more slowly as ΔF_0 increases, while the 50-ms data reproduces their dip at $\frac{1}{4}$ semitone ΔF_0, but shows some continued depression of identification rate at $\frac{1}{2}$ semitone. An analysis of variance was conducted, which covered the six ΔF_0's (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 semitones) and the two durations (50 ms/200 ms). The analysis showed significant main effects of ΔF_0 [$F(5,60)=12.89$, $p<0.0001$] and duration [$F(1,12)=31.21$, $p<0.0002$] and an interaction between ΔF_0 and duration [$F(5,60)=7.62$; $p<0.0001$].

The 200-ms stimuli gave significantly better identification than the 50-ms stimuli in the simple main effects at each nonzero ΔF_0 [$F(1)=7.03$, $p<0.05$; $F(1)=11.0$, $p<0.01$; $F(1)=5.8$, $p<0.05$; $F(1)=5.8$, $p<0.05$; $F(1)=8.4$, $p<0.02$]. Within the 200-ms condition, identification was significantly better in a Tukey (HSD) pairwise comparison at each nonzero ΔF_0 from that at zero ΔF_0 ($q=6.04$, 9.31, 8.83, 9.50, 12.09, respectively, $p<0.01$). The 4-semitone ΔF_0 condition was also significantly better than the $\frac{1}{4}$-semitone condition ($q=6.04$, $p<0.01$). Within the 50-ms condition, however, the only significant differences were produced by the dip at $\frac{1}{4}$ and $\frac{1}{2}$ semitones. The $\frac{1}{4}$-semitone identification rates were lower than in the 2- and 4-semitone conditions ($q=4.32$, $p<0.05$ and $q=5.18$, $p<0.01$) and the $\frac{1}{2}$-semitone identification rates were lower than in the 4-semitone condition ($q=4.32$, $p<0.05$).

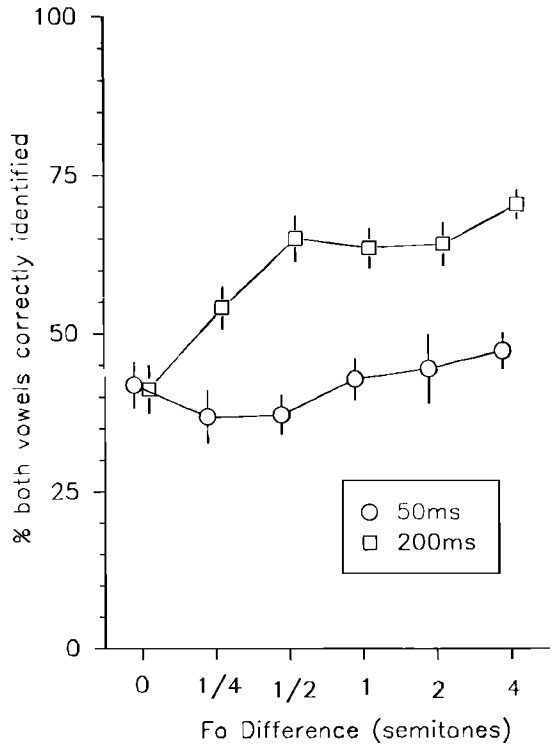

FIG. 2. Percent correctly identified vowel pairs and standard errors for subjects in experiment 1 for (1) 50 stimuli (circles) and (2) 200-ms stimuli (squares).

D. Discussion and conclusion

Our results generally replicate those of Assmann and Summerfield (1990), despite the use of only the exclusive set of vowel pairs and the use of a presentation level approximately 30 dB higher. Like them, we found that ΔF_0's improve identification for 200 ms but not for 50-ms double vowels. There are, however, two differences between their results and ours.

First, our overall scores are lower despite the use of only exclusive vowel combinations. This difference can probably be accounted for by the greater practice and experience of Assmann and Summerfield's subjects. Although our subjects gave variable overall response rates, all showed increased scores with increasing ΔF_0 for the 200-ms stimuli. However, in subsequent experiments we rejected subjects who score less than 55% overall.

Second, the drop in performance found here in the 50-ms data at $\frac{1}{4}$ and $\frac{1}{2}$ semitone ΔF_0 was only observed by Assmann and Summerfield in their $\frac{1}{4}$-semitone condition. They interpreted the drop as caused by the pattern of beats between corresponding *low* numbered harmonics from different vowels, which are not resolved separately by the peripheral auditory system. The overall beating pattern is cyclic, having a period equal to the difference in F_0. These beats may make the identity of the constituents of a double vowel more or less recognizable from the combined spectrum at different points in the cycle. The observed performance dip can be explained if the majority of the vowel pairs are relatively unrecognizable in the small portion of the cycle which was heard by the subjects in the 50 ms/$\frac{1}{4}$-semitone condition.

The dip in performance in our data at $\frac{1}{4}$ and $\frac{1}{2}$ semitone is consistent with this interpretation, since we used double vowels whose relative glottal phases were very similar to the relative phases used by Assmann and Summerfield (Summerfield, 1992), and since for $\frac{1}{4}$ and $\frac{1}{2}$ semitone ΔF_0's 50 ms is only 7% and 14% of the beating cycle, respectively. Presumably, the identities of the constituents become, on average, more recognizable from the combined spectrum in later parts of the cycle, allowing subjects' identification rates to recover at 1 semitone ΔF_0. The contribution of beating to the identification of double vowels will be discussed further in a forthcoming paper.

III. EXPERIMENT 2

A. Introduction

If, as Broadbent and Ladefoged (1957) suggested, a common F_0 allows the formants from a particular speaker to be grouped together appropriately, then identification of double vowels should be impaired when such grouping is disrupted. In this experiment across-formant grouping was disrupted by swapping the F_0's of the two vowels of a pair in a particular frequency region. The constituents of these "F_0-swapped" double vowels thus had an inconsistent F_0 across their formants; the first formant of one vowel was synthesised on the same F_0 as the higher formants of the other vowel, and vice versa (for the purposes of the experiments in this paper, the frequency region of the first formant is defined as extending up to the spectral minimum between $F1$ and $F2$ of the vowel in question). Across-formant grouping should be severely impaired by this manipulation, since it would lead to the wrong groupings of first and higher formants. On the other hand, this manipulation generally maintains a ΔF_0 between the vowels of a pair *within* a particular formant region. If across-formant grouping is responsible for the improvement in identification of double vowels found in experiment 1, then the improvement with increasing ΔF_0 will be severely reduced, or even reversed in the F_0-swapped condition of the following experiment.

B. Stimuli

Vowel duration was extended to 1000 ms, since other experiments had produced a more reliable improvement in identification with increasing ΔF_0 at this longer duration (Culling, 1991).

The first stage in the synthesis procedure produced 1000 ms "half-vowels" which contained either just the first formant or just the higher formant regions of each vowel. A cross-over frequency f_x was selected between the $F1$ and $F2$ of each vowel, near the minimum of its spectral envelope (/i/, 1250 Hz; /ɑ/, 800 Hz; /u/, 650 Hz; /ɔ/, 600 Hz; /ɜ/, 900 Hz). Each half-vowel was then synthesized with each of the six F_0's (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 semitones above 100 Hz) by digitally adding together only frequency components that were either lower than f_x-50 Hz, or higher than f_x+50 Hz. The resulting 100-Hz spectral notch centered on the cross-over frequency was used to eliminate the possibility of beating between components of the same vowel above and below the crossover. The presence of the notch had little influence on the phonetic quality of the vowels.

Half-vowels were then combined to produce full vowels which had either the same F_0 across the whole spectrum ("normal" vowels), or vowels with an abrupt change in F_0 between $F1$ and $F2$ ("split" vowels). The split vowels either stepped up from 100 Hz F_0 in the $F1$ region to a higher F_0 in the rest of the spectrum, or stepped down from the higher F_0 in the $F1$ region to 100 Hz in the rest of the spectrum. There were thus 30 normal vowels (6 F_0's$\times$5 vowels) and 60 split vowels (6 F_0's$\times$5 vowels$\times$2 step directions). It should be noted that the split vowels with no ΔF_0 were identical to the corresponding normal vowels, and were included only for statistical reasons.

Normal vowels were paired with other normal vowels to produce "normal" double vowels. Split vowels were paired with other split vowels, which used the same two F_0's, but in complementary spectral regions, to produce "F_0-swapped" double vowels. In such F_0-swapped double vowels, the lower harmonics of one vowel had the same F_0 as the higher harmonics of the other. Example spectra for the constituent vowels of an F_0-swapped pair are shown in Fig. 3. With 20 vowel combinations$\times$6 ΔF_0's, there were 120 normal and 120 F_0-swapped double vowels.

C. Procedure

Nine subjects (eight of whom had participated in at least one double-vowel experiment before) attended one hour-long session. After the practice test with the 30 individual normal vowels, subjects first received a further pretest using the 60 half-vowels, for which there was no performance criterion. Then subjects heard two tokens of each of the 240 normal and F_0-swapped double vowels in a random order which was changed for every third subject.

D. Results

Figure 4 shows the percentage of trials on which subjects correctly identified both members of a pair of normal or F_0-swapped vowels as a function of the ΔF_0. There is a

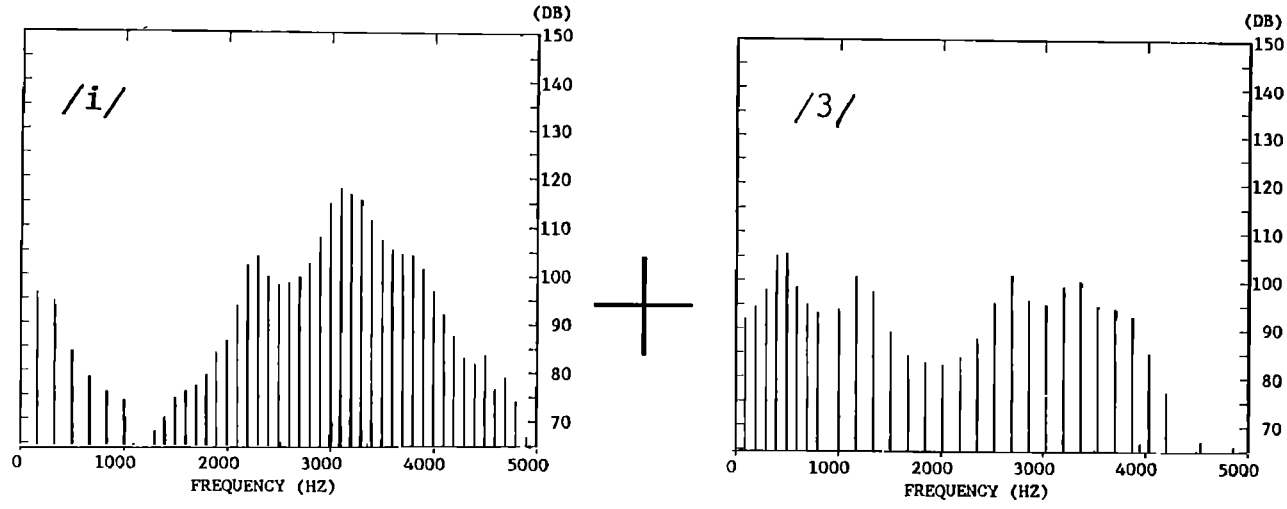

FIG. 3. Examples of the spectra of the constituents of an F_0-swapped stimulus: /i/+/3/ (for illustrative purposes an exaggerated ΔF_0 of 9 semitones is shown).

significant improvement in correct identification rate with increasing ΔF_0 [$F(5,40)=16.73$, $p<0.0001$], but no significant difference between the normal and F_0-swapped conditions, which are indistinguishable in the figure at ΔF_0's of $\frac{1}{4}$ and $\frac{1}{2}$ semitone. The F_0-swapped condition produced slightly poorer average identification rates than the normal condition at ΔF_0's of 1, 2, and 4 semitones. This trend is reflected in an interaction between ΔF_0 and stimulus type in the analysis of variance [$F(5,40)=3.08$, $p<0.02$], but the simple main effects did not show that the two conditions differed significantly at any ΔF_0 ($p>0.05$).

Confusion matrices for identification of the isolated half-vowels, presented in the half-vowel pretest, are given in Table II. Subjects correctly identified 43% of vowels from their $F1$ regions alone and 40% of vowels from the remaining formants alone.

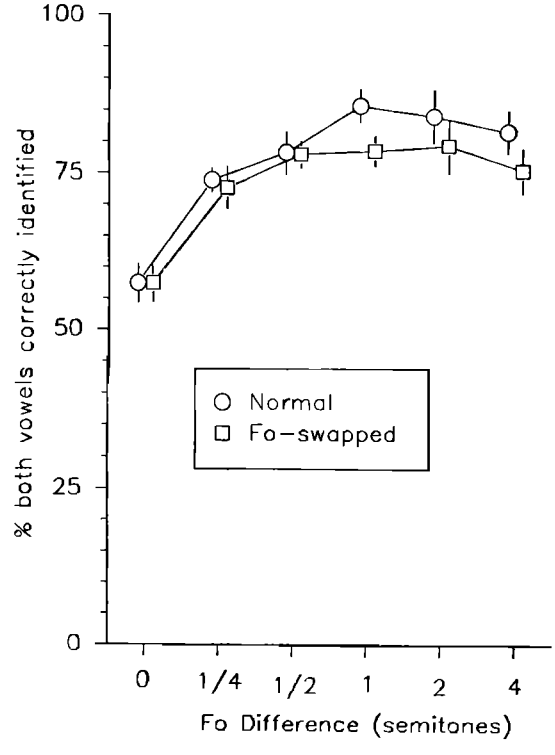

FIG. 4. Percent both vowels correctly identified in experiment 2 for (1) normal stimuli (circles) and (2) F_0-swapped stimuli (squares).

E. Discussion

If listeners group together the formants of a vowel by their common F_0, then they would group the $F1$ of one vowel of an F_0-swapped stimulus with the higher formants of the competing vowel. Such inappropriate grouping should impair their performance; but as Fig. 4 shows, the F_0-swapped double vowels showed only slightly worse performance than normal double vowels and then only with ΔF_0's of 1, 2, and 4 semitones. The improvement in identification for the F_0-swapped vowels with small ΔF_0's cannot be attributed to across-formant grouping. It remains possible that improved formant-frequency estimation, rather than across-formant grouping, is responsible for the improvement in identification at small ΔF_0's. This improved formant estimation could arise in the region of the first formant, the higher formants, or in both regions. Which frequency region contributes is addressed in experiment 4.

TABLE II. Confusion matrices showing percent responses of each class to each "half-vowel," for (a) $F1$ only (b) $F2$-5 only in the "half-vowel" pretest of experiment 2. Also, percent correct for combined stimuli, predicted from the "half-vowel" data using Boothroyd and Nittrouer [1988, Eq. (1)].

	Responses (%)										Prediction
	$F1$-only					$F2$-5					full vowel
Stim.	/i/	/ɑ/	/u/	/ɜ/	/ɔ/	/i/	/ɑ/	/u/	/ɜ/	/ɔ/	(% correct)
/i/	0	2	96	2	0	100	0	0	0	0	100
/ɑ/	0	37	0	59	4	0	96	0	0	4	98
/u/	22	0	76	2	0	0	56	9	30	5	78
/ɜ/	0	0	0	100	0	0	70	0	0	30	100
/ɔ/	0	0	33	67	0	0	11	2	74	13	13

The slightly lower performance for F_0-swapped stimuli relative to normal stimuli at ΔF_0's larger than half a semitone raises the possibility that this difference might continue to increase with larger ΔF_0's. This issue is addressed in experiment 3.

However, an alternative explanation for the results of experiment 2 must also be excluded. Although the "half-vowel" test shows that vowels are not easily identified from the upper or lower portions of their spectra alone, in the double-vowel experiments subjects may still have combined information from different frequency regions after the phonetic labeling stage. Had they done so their results would have been insensitive to the disruption of across-formant grouping cues. Boothroyd and Nittrouer [1988, Eq. (1)] relate the probabilities for identifying a token from either of two statistically independent sources of information in isolation (p_1 and p_2) to the probability of identification from both in combination (p_c):

$$p_c = 1-(1-p_1)(1-p_2). \quad (1)$$

This equation was applied to the probability of identifying a single vowel from its complete spectrum. Information from the first-formant region and from the region of the higher formants were taken to be statistically independent. The identification rates for the single complete vowels predicted under this assumption are included in Table II and show that most of the vowels could be identified well from independent phonetic categorization of the different frequency regions. The equation is somewhat generous, in that it assumes that new information always serves to improve the chances of a correct response rather than potentially misleading the listener. Nonetheless, as an additional control, experiment 3 also compared normal and F_0-swapped stimuli with those which contained only the $F1$ or only the $F2$-5 regions of the constituent vowels.

IV. EXPERIMENT 3

A. Introduction

The F_0-swapped stimuli in experiment 2 were designed to mislead mechanisms of across-formant grouping. Experiment 3 investigated two possible explanations for the failure to observe the substantial decline in performance predicted by across-formant grouping. First, across-formant grouping by F_0 may be weak only for the range of small ΔF_0's employed in experiment 2; and second, an individual spectral region may provide sufficient information for each vowel to be identified without integrating it with information from other spectral regions.

Accordingly, experiment 3 repeated the conditions of the previous experiment using normal and F_0-swapped double vowels, but extended the stimulus set to include larger ΔF_0's. As a control for vowel identification being based on information available only in a single frequency region, experiment 3 also asked subjects to identify vowel pairs consisting of only the first formant or only the higher formants of the constituent vowels: these are the "$F1$-only" and "$F2$-5" conditions, respectively.

B. Stimuli

Half-vowels were prepared with F_0's of 100, 102.29, 105.95, 112.25, 125.99, 141.42, 168.18, and 200 Hz. The F_0 of 102.29 was intended to be $\frac{1}{2}$ semitone above 100 Hz, but this value, entered in error (the correct F_0 being 102.93 Hz), was only 0.39 semitones above 100 Hz. The programmed nature of the stimulus preparation ensured that this error was consistent across all stimuli in the $\frac{1}{2}$-semitone condition. There were thus 8 ΔF_0 conditions altogether 0, "$\frac{1}{2}$," 1, 2, 4, 6, 9, and 12 semitones.

The normal and F_0-swapped stimulus types were prepared in the same way as in experiment 2 by adding together four half-vowels. The $F1$-only stimuli were prepared by adding together only the two $F1$ half-vowels. The $F2$-5 stimuli were prepared by combining only the two $F2$-5 half-vowels.

Since only two half-vowels composed the $F1$-only and $F2$-5 stimuli, these stimuli were lower in intensity than the normal and F_0-swapped stimuli, which were each composed of four half-vowels. In addition, due to the -6 dB/oct. spectral tilt on the Klatt synthesiser's combined voice source and radiation functions, the $F2$-5 half-vowels were considerably less intense than the $F1$ half-vowels, making the $F1$-only stimuli more intense than the $F2$-5 stimuli. No attempt was made to compensate for these intensity differences. With 8 ΔF_0's $\times 4$ conditions $\times 20$ vowel combinations, there were 640 stimuli altogether.

C. Procedure

Eight subjects, all of whom were experienced in double-vowel experiments, attended one hour-long session. The practice contained the 40 individual vowels which would compose the normal stimuli in the experiment to follow. The 640 stimuli were presented once in a random order which was changed for every third subject.

D. Results

Figure 5 shows that conditions which have full spectra (normal and F_0-swapped) give markedly better identification rates than those which have only a portion of the spectrum ($F1$-only and $F2$-5). This result was reflected in an analysis of variance covering the normal, F_0-swapped, $F1$-only, and $F2$-5 stimulus types and the 8 ΔF_0's (0, "$\frac{1}{2}$," 1, 2, 4, 6, 9, and 12 semitones) by a main effect of stimulus type [$F(3,21)=138.2$, $p<0.0001$]. Better identification of the normal and F_0-swapped stimuli was confirmed by Tukey pairwise comparisons; $F1$-only and $F2$-5 conditions differed significantly from each other, and each gave significantly lower identification rates than either the normal or F_0-swapped conditions ($p<0.01$ in each case). The normal and F_0-swapped conditions did not differ significantly ($q=3.09$, $p>0.05$).

Figure 5 also shows that identification rates improved with ΔF_0 only in the normal and F_0-swapped conditions. This interpretation was corroborated by the analysis of variance which showed a main effect of ΔF_0 [$F(7,49)=3.36$, $p<0.01$] and a significant interaction between stimulus type and ΔF_0 [$F(21,147)=4.20$, $p<0.0001$]. The sim-

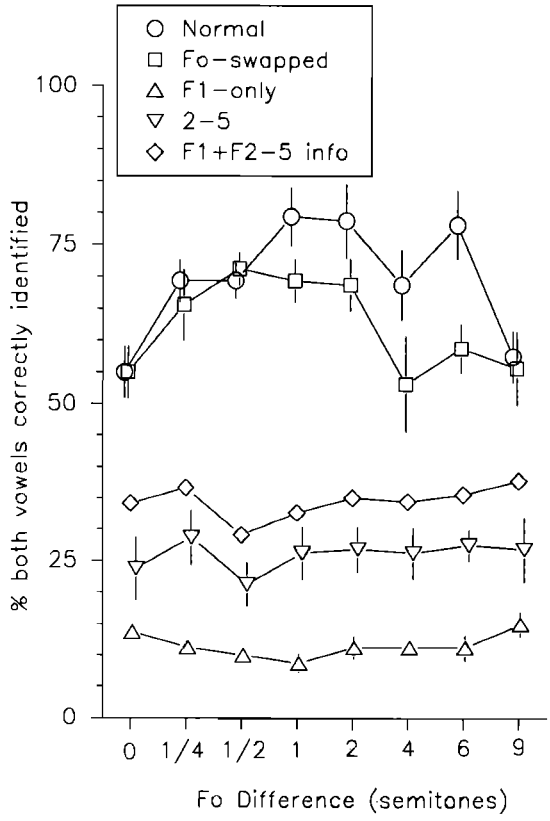

FIG. 5. Percent both vowels correctly identified in experiment 3 for (1) Normal stimuli (circles), (2) F_0-swapped stimuli (squares), (3) $F2$-5 stimuli (upright triangles), (4) $F1$-only stimuli (inverted triangles), and (5) predicted from $F1$-only and $F2$-5 scores (diamonds).

ple main effects confirmed that the effect of ΔF_0 was significant only in the normal and F_0-swapped conditions [$F(7)=699.7$, $p<0.0001$ and $F(7)=431.3$, $p<0.005$, respectively]. Although the figure shows that the normal stimuli gave higher identification rates than the F_0-swapped stimuli at ΔF_0's of 2–9 semitones, Tukey pairwise comparisons at each ΔF_0 showed a significant difference between these conditions only at 9 semitones ΔF_0 ($q=4.64$, $p<0.05$).

E. Discussion and conclusions

First we consider the identification scores for the control conditions. What level of identification would we expect in the normal and the F_0-swapped conditions if listeners were simply basing their judgements on pairs of phonetic labels derived independently from the two different formant regions? Again, taking information from the first-formant region and the higher formant regions as statistically independent, Eq. (1) from Boothroyd and Nittrouer (1988) can be used to predict identification rates for complete double vowels from scores in the $F1$-only and $F2$-5 conditions. The results of this analysis are shown as the stars in Fig. 5. The predictions not only fall far short of the level of performance found in the normal and in the F_0-swapped conditions but also do not show any improvement with increasing ΔF_0. Apparently, subjects did not identify the vowels by independently labeling separate spectral regions.

We now turn to the difference in identification between the normal and the F_0-swapped conditions. The significantly impaired performance for F_0-swapped as compared to normal stimuli at a ΔF_0 of 9 semitones shows that the effect of misleading across-formant grouping mechanisms in the F_0-swapped condition is stronger at higher ΔF_0's. Across-formant grouping mechanisms could be responsible for this clear deterioration in vowel identification in the F_0-swapped condition.

Why should across-formant grouping mechanisms exert their effect only at relatively large ΔF_0's? For most of the vowels used here, the frequency components in the second-and-higher-formant region will not be resolved. In order for listeners to use a ΔF_0 to segregate formants, they would have to be able to detect two different fundamental frequencies in each frequency region and correctly match up F_0's which were common to different frequency regions. Listeners can detect differences in F_0 across regions of resolved and unresolved harmonics (Carlyon *et al.*, 1992), but relatively large differences in F_0 are required (1–2 semitones). This result is undoubtedly related to the fact that the fundamental frequency difference limen for complex periodic sounds consisting of only unresolved harmonics is about an order of magnitude greater than that for sounds that contain some resolved harmonics (Houtsma and Smurzynski, 1990).

The large ΔF_0 of 9 semitones required to produce a significant difference between the normal and the F_0-swapped conditions in the present experiment can be compared with previous results by Gardner *et al.* (1989). They found that a ΔF_0 of about 4.4 semitones was required to produce perceptual separation of $F2$ from the other simultaneously presented formants of a composite ru/li syllable.

V. EXPERIMENT 4

A. Introduction

The primary goal of this experiment was to identify the frequency region responsible for the improved identification with small ΔF_0's in the F_0-swapped double vowels of the previous two experiments. The experiments by Houtsma and Smurzynski referred to above showed that listeners are more sensitive to sequentially presented F_0 differences between resolved than between unresolved harmonics. If listeners are relatively insensitive to small differences in F_0 between sounds consisting of only unresolved harmonics, then the improvement in identification scores that we have found at small ΔF_0's with the F_0-swapped stimuli could be due entirely to F_0 differences in the first formant region near the $F1$ peak, where the harmonics are resolved.

To test this hypothesis we produced another type of double-vowel stimulus in which the difference in F_0 occurred only in first formant region; both the higher-formant regions had the same F_0. These "same$F2$-5" double vowels should be identified as accurately as the normal

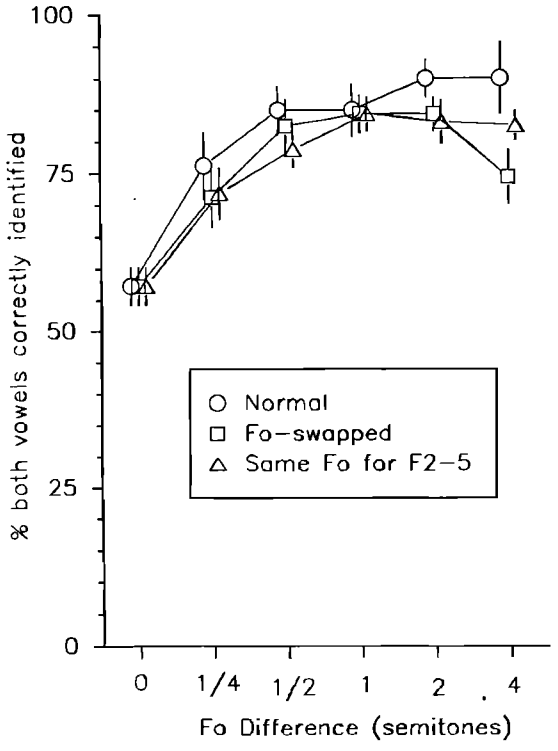

FIG. 6. Percent both vowels correctly identified in experiment 4 for (1) Normal stimuli (circles), (2) F_0-swapped stimuli (squares), and (3) same$F2$-5 stimuli (triangles).

and the F_0-swapped double vowels if listeners are insensitive to small ΔF_0's in the higher formants.

B. Stimuli

The half-vowels from experiment 2 were used again in this experiment in order to recreate the normal and F_0-swapped conditions, plus the new same$F2$-5 condition. This condition was made by combining the half-vowels in a different way. As before, the same$F2$-5 stimuli were composed from $F1$ half-vowels with different F_0's, but for this condition the $F2$-5 half-vowels both had the same F_0, so that only one F_0 was present in this frequency region. Two versions of each same$F2$-5 stimulus were made, which differed in which of the two F_0's was used for the $F2$-5 region. With 6 ΔF_0's (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 semitones) $\times$ 20 vowel combinations, this design resulted in 120 normal, 120 F_0-swapped, and 240 same$F2$-5 double vowels.

C. Procedure

Seven subjects from previous experiments and one who was inexperienced in double-vowel experiments attended one hour-long session. The 480 experimental stimuli were presented once to each subject in a random sequence, which was changed for every second subject.

D. Results

Figure 6 shows that all three conditions yielded similar improvements in identification rates as ΔF_0 increased from 0 to 1 semitone. An analysis of variance compared the normal, F_0-swapped, and same$F2$-5 conditions across the 6 ΔF_0's (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 semitones). This analysis revealed significant main effects of stimulus type [$F(2,7) = 10.1$, $p < 0.002$] and ΔF_0 [$F(5,35) = 18.6$, $p < 0.0001$], but no interaction. Overall, the accuracy of identification for F_0-swapped and same$F2$-5 stimuli was somewhat lower than for normal stimuli and the conditions begin to diverge more sharply at 4 semitones. Although this divergence did not produce an interaction in the ANOVA, the simple main effects showed that there were significant differences between the three stimulus types only at 4 semitones ΔF_0 [$F(2) = 7.18$, $p < 0.01$]. All three stimulus types produced significant simple main effects of ΔF_0 [$F(5) = 5.58$, $p < 0.001$; $F(5) = 6.53$, $p < 0.0005$; $F(5) = 4.62$, $p < 0.005$].

E. Discussion

Identification of the double vowels increased over the first semitone of ΔF_0 by about the same amount in all three conditions of this experiment. In particular, the same$F2$-5 condition shows as large an increase as the F_0-swapped condition. This increase in identification cannot then be due to mechanisms operating in the higher-formant region (including across-formant grouping) since, in this region, each vowel in the same$F2$-5 condition had the same F_0 as its partner. The increase must be due to mechanisms operating solely in the first formant region.

With a 4 semitone ΔF_0 there is again some suggestion that mechanisms concerned with the higher-formant region can influence identification. The F_0-swapped and the same$F2$-5 conditions were both slightly worse than the normal condition.

VI. EXPERIMENT 5

A. Introduction

Experiment 4 showed that ΔF_0's in the $F1$ region are sufficient to produce the improvement in identification accuracy for double vowels with small ΔF_0's. One explanation for this effect is offered by physiological data. Miller and Sachs (1984) found that the neural representation of within-channel amplitude modulation, thought to encode pitch for high numbered harmonics (Schouten, 1940), cannot be detected in auditory nerve microelectrode recordings at high presentation levels.

Experiments 1–4 used stimuli at levels around 85–90 dB(A). At these high levels, perceptual separation in the $F2$-5 region may, therefore, be lost through impaired encoding of amplitude modulation. In contrast, Scheffers (1983) presented his stimuli at around 60 dB (SPL), Zwicker presented his stimuli at 63 dB(A), Assmann and Summerfield used only 53 dB(A), while Chalikia and Bregman used 65 dB(A). In some respects the higher presentation levels used here represent a more realistic model of the cocktail party effect, which is intrinsically concerned with the auditory system's response to noisy listening environments, rather than the 65–70 dB levels typical of speech in a quiet listening environment.

Experiment 5 was designed to assess the influence of presentation level on across formant grouping by F_0. Ex-

periment 5 replicated experiment 2 using two presentation levels: 85–90 dB(A), as used in experiments 1–4, and 55–60 dB(A), typical of previous research. If the encoding of amplitude modulation, and hence pitch, has been impaired at high presentation levels then a lower level will improve $F2$-5 segregation by ΔF_0; at the lower level, performance with F_0-swapped stimuli should decline (relative to that with normal stimuli) at smaller ΔF_0's than at the higher level.

B. Stimuli

The stimuli were identical to those of experiment 2, with normal and F_0-swapped stimulus types and 6 ΔF_0's (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 semitones) giving 240 stimuli. Attenuation of 30 dB in the low intensity condition was achieved by inserting separate Advance Electronics A64 step attenuators into each ear's signal channel after digital-to-analog conversion.

C. Procedure

Eight subjects, experienced in double-vowel experiments, attended one hour-long session, which was divided into two blocks with different presentation levels. Each stimulus was presented once in each block. Four subjects received a block with the higher intensity first, while the other four started with the lower intensity block. Two different random sequences were used across subjects.

D. Results

Figure 7 shows that the overall pattern of results is similar to those of experiment 2; in each condition identification improves over the first semitone and decreases somewhat for the F_0-swapped conditions at 4 semitones. These results were reflected in an analysis of variance covering stimulus type (normal and F_0-swapped), stimulus level (quiet and loud), and ΔF_0 (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4 semitones) by a main effect of ΔF_0 [$F(5,35)=11.5$, $p<0.0001$] and an interaction between stimulus type and ΔF_0 [$F(5,35)=6.1$, $p<0.0005$].

Identification was better for the loud presentation level [$F(1,7)=15.7$, $p<0.01$] and better for normal than for F_0-swapped stimuli [$F(1,7)=8.1$, $p<0.05$]. The figure appears to show a dip in correct identification rate at the loud presentation level for a ΔF_0 of $\frac{1}{2}$ semitone and also that there is some difference between the normal and F_0-swapped conditions at 2 semitones using the quiet level. There was, however, no significant change with level in the pattern of improvement with ΔF_0, as reflected in a three-way interaction between stimulus type, level, and ΔF_0 [$F(5,30)=2.10$, $p>0.05$].

E. Discussion and conclusions

The results of experiment 5 are consistent at each presentation level with the results of experiments 2 and 4, in that normal and F_0-swapped stimuli give similar improvements in identification at low ΔF_0's. With regard to the specific hypothesis that pitch mechanisms operating in the higher frequency regions may be more effective at lower levels, this data offers only limited support in the form of an appropriate but small and nonsignificant trend (the difference between the normal and F_0-swapped conditions at only 2 semitones ΔF_0 at the lower presentation level). The results are consistent with the hypothesis, but indicate that any effect is quite small. The results from this experiment do not, therefore, detract from the general finding that pitch mechanisms based on unresolved harmonics can play only a minor role in the improvement in double-vowel identification with ΔF_0's of a semitone or less.

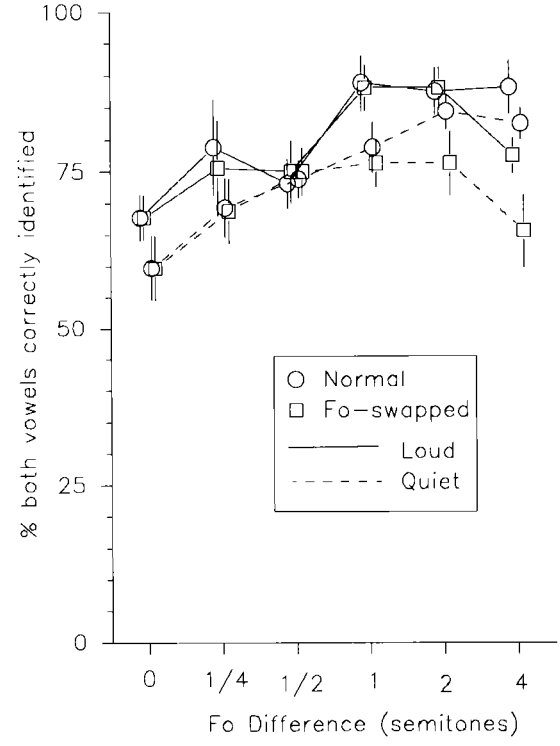

FIG. 7. Percent both vowels correctly identified in experiment 5 for (1) loud and normal (squares, solid line), (2) loud and F_0-swapped (circles, solid line), (3) quiet and normal (squares, dashed line), and (4) quiet and F_0-swapped (circles, dashed line).

Since stimulus level has little effect on across-formant grouping, the ineffectiveness of across-frequency grouping at small ΔF_0's has some generality. It remains, therefore, to explore the conditions in which a role can be demonstrated for ΔF_0's in the higher formant region.

Since the results of experiment 5 showed that the higher stimulus level used hitherto improves listeners' accuracy of identification, but does not have different effects in different conditions, further experiments continued to use this high level.

VII. EXPERIMENT 6

A. Introduction

Experiment 6 examined further the contribution of ΔF_0's in the higher formant region to double-vowel identification at the larger ΔF_0's used in experiment 3. The normal, F_0-swapped, and same$F2$-5 conditions from experiment 4 were extended to an octave range of ΔF_0's. Exper-

iment 4 showed that across-formant inconsistencies in F_0 impaired identification of F_0-swapped double-vowels at ΔF_0's greater than about 2 semitones. On the basis of this finding the same$F2$-5 condition should also be worse than the normal condition at larger ΔF_0's, since it also will be susceptible to the effect of across-formant inconsistencies in F_0.

A new stimulus condition was introduced, same$F1$, in which vowel pairs had different F_0's only in the higher formant region. In this condition the ΔF_0 in the higher formant region must be responsible for any observed improvement in identification with increasing ΔF_0. On the basis of previous experiments, we predicted that any improvement in identification with increasing ΔF_0 in this condition should be slight. Small ΔF_0's should be ineffective in the higher formant region because they are close to the difference limen for fundamental frequency for unresolved harmonics, while at higher ΔF_0's grouping of the first formant with higher formants should be disrupted by an inconsistency between the F_0's in the first and the higher formant regions.

B. Stimuli

A new set of "half-vowels" was synthesized with F_0's of 100, 105.95, 112.46, 125.99, 141.42, 168.18, and 200 Hz, using the same cross-over frequencies as experiment 2. The higher F_0's of the stimuli were thus 0, 1, 2, 4, 6, 9, or 12 equal-tempered semitones above 100 Hz.

These half-vowels were combined in order to produce the normal, F_0-swapped, and same$F2$-5 conditions used in experiment 4, plus a fourth condition, same$F1$, for which the $F1$ half-vowels were both at the same F_0, while the F_0's of the higher formants differed. As in experiment 4 there were two versions of each same$F2$-5 pair at each ΔF_0, one with the higher formant region at 100 Hz F_0, and a second using the other F_0. Similarly, two versions were also used for the same$F1$ condition, so that, with 20 vowel combinations $\times 7$ ΔF_0's, there were 140 normal stimuli, 140 F_0-swapped stimuli, 280 same$F2$-5 stimuli, and 280 same$F1$ stimuli.

C. Procedure

Nine subjects (of whom seven had performed experiments 2 and/or 4 and two were inexperienced in double-vowel experiments) attended two hour-long sessions. Before each session subjects identified all the 35 individual vowels which would constitute the normal stimuli in the experiment.

Due to extensions of the experiment introduced after some data had already been collected, counterbalancing was incomplete. The two sessions consisted of different sets of conditions: session A contained the normal stimuli, the F_0-swapped stimuli, and the same$F2$-5 stimuli; session B contained the normal stimuli, the F_0-swapped stimuli and the same$F1$ stimuli. Each stimulus from each of these conditions was presented once per session. Seven of the subjects had already completed session A before session B was added to the experiment. A further two subjects did session B first, and then session A. Overall each subject received 280 stimuli from each condition: 280 same$F2$-5 stimuli in session A, 280 same$F1$ stimuli in session B, and two presentations (in different sessions) of the 240 normal and 240 F_0-swapped stimuli.

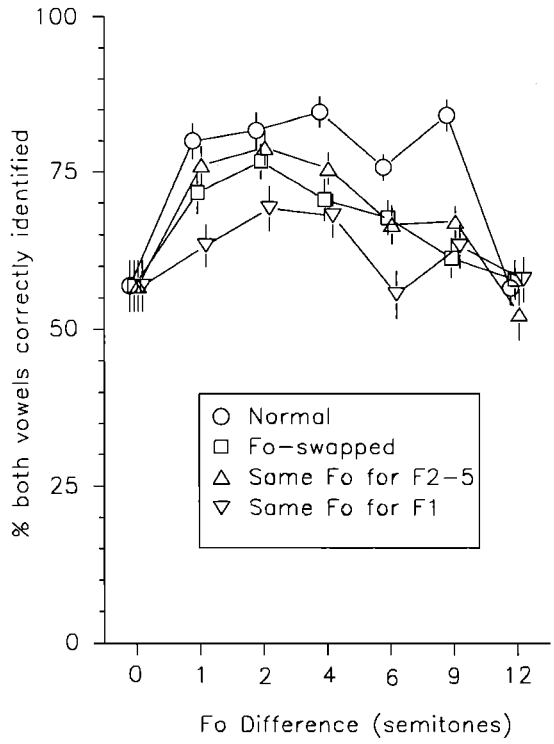

FIG. 8. Percent both vowels correctly identified in experiment 6 for (1) normal stimuli (circles), (2) F_0-swapped stimuli (squares), (3) same$F2$-5 (upright triangles), and (4) same$F1$ (inverted triangles).

D. Results

The overall pattern of results (Fig. 8) is consistent with previous experiments. Accuracy of identification improves steeply with a ΔF_0 of one semitone in the normal, the F_0-swapped, and the same$F2$-5 conditions. The latter two conditions show a decline in identification accuracy at higher ΔF_0's. The effect of ΔF_0's in the same$F1$ condition is smaller and grows more slowly with increasing ΔF_0, before declining rapidly at 6 semitones ΔF_0. All four conditions show similar identification accuracy with one octave (12 semitones) ΔF_0 as at zero ΔF_0.

An analysis of variance was conducted covering the four stimulus types (normal, F_0-swapped, same$F2$-5, and same$F1$), two repeated measures (repeated presentations of normal and F_0-swapped stimuli; different versions of same$F2$-5 and same$F1$ stimuli) and 7 ΔF_0's (0, 1, 2, 4, 6, 9, and 12 semitones). The analysis revealed significant differences between the four stimulus types [$F(3,24)=13.9$, $p<0.0001$] and between the seven ΔF_0's [$F(6,48)=18$, 0, $p<0.0001$]. There were also interactions between these two factors [$F(18,144)=3.75$, $p<0.0001$], between repeated measures and ΔF_0 [$F(6,48)=2.7$, $p<0.05$] and between all three factors [$F(18,144)=2.93$, $p<0.0005$].

The simple main effects showed that the effect of ΔF_0 was highly significant for the normal, F_0-swapped, and

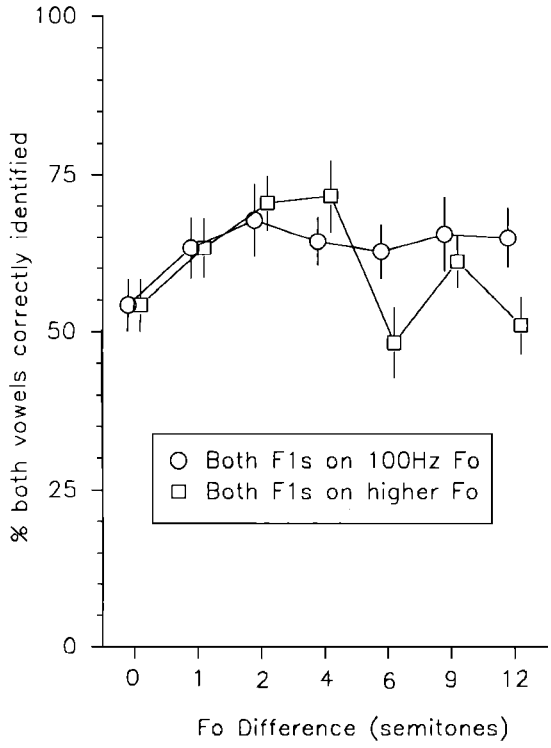

FIG. 9. Percent both vowels correctly identified in experiment 6 for same$F1$ stimuli, the $F1$'s of which were both excited by (1) 100 Hz F_0 (circles) and (2) the higher F_0 (squares).

same$F2$-5 conditions [$F(6)=9.39$, $p<0.0001$; $F(6)=7.37$, $p<0.001$; $F(6)=5.23$, $p<0.0005$], but was only just significant in the same$F1$ condition [$F(6)=2.34$, $p<0.05$], reflecting the smaller increase in correct identification with ΔF_0 in this condition. The differences between the four stimulus types were significant at 1, 4, 6, and 9 semitones ΔF_0 [$F(3)=3.5$, $p<0.05$; $F(3)=3.7$, $p<0.05$; $F(3)=4.78$, $p<0.01$; $F(3)=7.4$, $p<0.002$], reflecting superior identification accuracy for normal double vowels compared to one or more of the other conditions. A Tukey pairwise comparison between the four stimulus types showed that the normal condition gave significantly more accurate identification that the same$F1$ condition at 1, 4, 6, and 9 semitones ΔF_0 ($q=4.37$, 4.37, 5.32, and 5.46, respectively). The nonsignificance of the simple main effect and the normal–same$F1$ pairwise comparison at 2 semitones ΔF_0 reflects the relatively good recognition accuracy which subjects achieved with same$F1$ stimuli at that ΔF_0. The pairwise comparisons also showed that the normal condition was significantly better than the F_0-swapped and same$F2$-5 conditions at 9 semitones ΔF_0 ($q=4.44$, $p<0.05$; $q=5.97$, $p<0.01$).

All four conditions show a decline in performance between 4 and 6 semitones ΔF_0. For the normal condition performance recovers at 9 semitones, forming a dip in recognition accuracy at 6 semitones ΔF_0. For the F_0-swapped and same$F2$-5 conditions it forms part of a progressive decline which begins at 4 semitones ΔF_0, while in the same$F1$ condition identification accuracy collapses at this point to zero ΔF_0 levels. Figure 9 shows that the latter drop in accuracy is due mainly to those versions of the stimuli in which the first formant region is excited by the higher of the two F_0's. This effect of same$F1$ version is probably responsible for the interactions between repeated measures (version) and ΔF_0 and between stimulus type, repeated measures and ΔF_0.

E. Discussion

1. Effects of across-formant grouping

The use of across-formant grouping in the identification of double-vowel stimuli is demonstrated in experiment 6 by the significantly lower identification scores for the F_0-swapped condition compared to the normal condition at 9 semitones ΔF_0 (see Fig. 8). Performance is depressed in the F_0-swapped condition because across formant grouping mechanisms are being misled.

One of the constituent vowels in a same$F2$-5 stimulus pair also contains an across-formant inconsistency in F_0, which may upset across-formant grouping. The identification rates for same$F2$-5 stimuli follow closely the trends shown by the F_0-swapped data across all ΔF_0's and are also significantly poorer than the normal data at 9 semitones ΔF_0. This outcome suggests that across-formant grouping mechanisms are being confused by same$F2$-5 stimuli in the same way as by F_0-swapped stimuli.

2. Improvement in F2 and F3 separation

In the same$F1$ condition there is a ΔF_0 only between the higher formants of the vowels. Given that improvement in identification accuracy is mediated by separation processes, it is clear from the results of the same$F1$ condition (see Figs. 8 and 9) that separation processes which exploit ΔF_0's are active in the $F2$-5 region. In addition, the effects of across-formant grouping, discussed above, are indirect evidence for such processes. The maximum same$F1$ performance occurs at 2–4 semitones ΔF_0. In comparison, the much bigger effect mediated by the $F1$ region (observed here and in experiment 4 through the same$F2$-5 stimuli), is complete after the introduction of only 1 semitone ΔF_0. Thus larger ΔF_0's are required in order to separate the higher formants than to separate the first formants. The need for larger ΔF_0's for separation in higher frequency regions accords with the results of Gardner *et al.* (1989) who found that the ΔF_0 required to detect the presence of a second source (with around 90% reliability) was greater for a mistuned fourth formant ($\cong$4.4 semitones) than for a mistuned second formant ($\cong$0.6 semitones).

On the other hand, comparison of the F_0-swapped and same$F2$-5 conditions does not support a role for ΔF_0's in the $F2$-5 region. Figure 8 shows no trend for the F_0-swapped stimuli, which have ΔF_0's in the $F2$-5 region, to give progressively superior scores to the same$F2$-5 stimuli, which do not. Indeed, averaged across ΔF_0's the $F0$-swapped stimuli tend to give lower scores.

The apparently conflicting evidence for and against separation in the $F2$-5 region can be resolved if one considers the relative dominance of the $F1$ region. In the F_0-swapped and same$F2$-5 conditions the ΔF_0 in the $F1$ re-

gion has a powerful effect, raising performance markedly. Any small contribution from improved $F2$ and $F3$ frequency estimation may be swamped by the $F1$ effect.

It should be noted that a separation effect observed in the $F2$-5 region does not necessarily imply an effect mediated by unresolved harmonics, since for some vowels the second formant was at least partially excited by peripherally resolvable harmonics. A recent extension of the "ru/li" paradigm (Darwin, 1992), in which the second formant was mistuned in F_0 from syllables synthesized with F_0's of 80, 120, or 200 Hz, suggests that perceptual exclusion of that formant (and hence perception of /li/, rather than /ru/) is more dependent upon the absolute F_0 of the formant than of the ΔF_0. The F_0 at which a categorical transition occurred was consistent with the hypothesis that the formant had to be excited by resolved harmonics for perceptual exclusion of that formant to occur.

3. Other effects

An unexpected issue arose in the same$F1$ condition. Here each vowel combination was represented by two stimuli at each ΔF_0; for one, the $F1$ region of both vowels was excited by 100 Hz F_0, while for the other, both $F1$'s were excited by the higher F_0. In the same$F1$ condition, there was a drop in performance for stimuli which had the $F1$'s of both vowels on an F_0 of 100 Hz + 6 semitones or 141 Hz (see Fig. 9). It seems likely that this drop is caused by a poor definition of the combined $F1$ spectral envelope. Similar drops in performance were observed in the other three conditions as well as in the 6 semitone ΔF_0 conditions of experiment 3 (for normal and F_0-swapped stimuli). The poor definition of the $F1$ envelope appears to be related to the particular harmonic spacing of 141 Hz, since two sources of evidence show that other large harmonic spacings do not have the same effect. First, in both experiments 3 and 6, there is some recovery in identification accuracy at 9 semitones ΔF_0, before the drop at one octave, and second, Assmann (1992) has found that the identification accuracy for double-vowels which have a baseline F_0 of 200 Hz is very similar to that for double vowels which have a 100-Hz baseline. These results show that widely spaced harmonics do not in themselves disrupt listeners' identification accuracy.

VIII. GENERAL DISCUSSION

Experiments with simultaneous ("double") vowels show that differences in fundamental frequency (ΔF_0's) as small as $\frac{1}{4}$ semitone (1.5%) make the two vowels much easier to identify (Scheffers, 1983; Zwicker, 1984; Chalikia and Bregman, 1989; Assmann and Summerfield, 1990). This effect has been attributed to mechanisms which group parts of the sound which have the same F_0 and segregate parts which have different F_0's. However, the limited number of experiments which have demonstrated that the auditory system segregates the formants of speech sounds on the basis of common F_0 have required much larger ΔF_0's (Darwin, 1981; Gardner *et al.*, 1989). The present study investigated the mismatch between the size of ΔF_0 needed to segregate formants and that needed to improve listners' identification accuracy for double vowels by using constituent vowels with across-formant inconsistencies in F_0 in double-vowel experiments. These inconsistencies were designed to confuse any grouping/segregation of formants according to their F_0's.

Experiments 2–6 compared the accuracy of identification for stimuli composed from vowels with and without across-formant inconsistencies in F_0 (the normal and F_0-swapped conditions). In either case, performance increased markedly with increasing ΔF_0 up to 1 semitone, showing that this increase cannot be attributable to across-formant grouping mechanisms, which ought to be confused by the inconsistent F_0's of the constituent vowels. ΔF_0's of at least 4 semitones were required before an across-formant grouping effect, in the form of slightly lower performance in the conditions with inconsistent F_0's, was large enough to be significant, although the normal condition always yielded slightly higher scores at even the smallest ΔF_0's.

Experiments 3 and 6 showed that when larger ΔF_0's of 6 and 9 semitones are used, the inconsistent F_0's can disrupt performance more, indicating that, in line with the results of Darwin and Gardner *et al.*, across-formant grouping/separation mechanisms require large ΔF_0's. Experiment 3 also showed that the double vowels were not recognizable using either the first-formant region or the higher formant region alone, and that identification based either on these isolated regions, or upon the phonetic labels derived independently from both regions, does not improve with increasing ΔF_0. So, although across-formant grouping by F_0 has only a minor role in double-vowel experiments, information from the first-formant region and the higher formant region must be integrated in some way prior to phonetic categorization.

Experiment 5 showed that the role played by across-formant grouping was not significantly altered when the presentation level was reduced from around 90 to 60 dB(A), but that overall performance was significantly poorer with the lower presentation level.

The powerful effect of small ΔF_0's on the identification of double vowels was investigated in experiments 4 and 6. Since across-formant grouping by F_0 has little effect upon identification rates for low ΔF_0's, it was possible to construct stimuli which possessed ΔF_0's only in the first-formant region or only in the higher formant region, without inappropriate formant grouping disrupting performance. These experiments showed that ΔF_0's in the first-formant region were required for large improvements in identification, and that ΔF_0's in the higher formant region produced smaller effects, which required larger ΔF_0's. So, the large effect of small ΔF_0's in double-vowel experiments must be attributed chiefly to mechanisms operating in the first-formant region. This pattern of results is compatible with Houtsma and Smurzynski's data on pitch difference limens (Houtsma and Smurzynski, 1990). They found that the pitch difference limen grew sharply when the stimuli contained only harmonics above the 10th, which tend not to be resolved by the peripheral auditory system. Listeners' ability to identify the F_0 which excites a

particular frequency region of a sound is, therefore, highly dependent on the harmonic number of the harmonics in that region, and in a double-vowel experiment the higher formants are much more likely to be excited by high numbered harmonics than the first formant. Hence, when the ΔF_0 is small, listeners may be able to determine the two F_0's in the first-formant region, and so group components of common F_0 within that region, but they are poorer at determining the F_0's which excite the higher formants. When the F_0's in the higher formant region are not determined, listeners are able neither to group the components which excite the higher formants, nor to group the formants themselves on the basis of common F_0.

IX. CONCLUSIONS

The results of these six experiments show a fairly consistent overall pattern.

(1) At even the smallest ΔF_0's used ($\frac{1}{4}$ semitone) there is a powerful effect on identification accuracy mediated by the $F1$ region, which accounts for most of the 25% increase in performance over the first semitone of ΔF_0.

(2) There is little evidence effects of ΔF_0's in the $F2$-5 region at the smallest ΔF_0's. Firmer evidence of its smaller contribution is only visible at ΔF_0's of 2–4 semitones.

(3) Across-formant grouping effects also show little sign of emerging at low ΔF_0's, regardless of presentation level, and only emerge strongly for ΔF_0's of at least four semitones.

ACKNOWLEDGMENTS

We wish to thank Quentin Summerfield, David Bailey, and Ray Meddis for reading and providing valuable comment on various versions of this manuscript. The SERC provided the laboratory with equipment on various grants and supported John Culling with a research studentship.

Assmann, P. F. (**1992**). Personal communication.

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.

Broadbent, D. E., and Ladefoged, P. (**1957**). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. **29**, 708–710.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon. **10**, 23–36.

Carlyon, R. P., Demany, L., and Semal, C. (**1992**). "Detection of across-frequency differences in fundamental frequency," J. Acoust. Soc. Am. **91**, 279–292.

Chalikia, M. H., and Bregman, A. S. (**1989**). "The perceptual separation of simultaneous auditory signals: Pulse train segregation and vowel segregation," Percept. Psychophys. **46**, 487–496.

Culling, J. F. (**1991**). "The perceptual separation of concurrent vowels," Ph.D. thesis, Sussex University.

Cutting, J. E. (**1976**). "Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening," Psychol. Rev. **83**, 114–140.

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset-time," Q. J. Exp. Psychol. A **33**, 185–207.

Darwin, C. J. (**1992**). "Listening to two things at once," in *The Auditory Processing of Speech*, edited by M. E. H. Schouten (Mouton & Gruyer, Berlin).

Gardner, R. B., Gaskill, S. A., and Darwin, C. J. (**1989**). "Perceptual grouping of formants with static and dynamic differences in fundamental frequency," J. Acoust. Soc. Am. **85**, 1329–1337.

Houtsma, A. J. M., and Smurzynski, J. (**1990**). "Pitch identification and discrimination for complex tones with many harmonics," J. Acoust. Soc. Am. **87**, 304–310.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 838–844.

Meddis, R., and Hewitt, M. J. (**1992**). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233–245.

Miller, I. M., and Sachs, M. B. (**1984**). "Representation of voice pitch in discharge patterns of auditory-nerve fibers," Hear. Res. **14**, 257–279.

Moore, B. J. C., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

Parsons, T. W. (**1976**). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. **60**, 911–918.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.

Scheffers, M. T. M. (**1982**). "The role of pitch in perceptual separation of simultaneous vowels II," IPO Ann. Prog. Rep. **17**, 41–45.

Scheffers, M. T. M. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, Gronigen.

Schouten, J. F. (**1940**). "The residue and the mechanism of hearing," Proc. Kon. Ned. Akad. Wetensch. **41**, 1086–1093.

Stubbs, R. J., and Summerfield, Q. (**1990**). "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **87**, 359–372.

Summerfield, Q., and Assmann, P. F. (**1991**). "Perceptual separation of concurrent vowels: Effects of pitch pulse asynchrony and harmonic misalignment," J. Acoust. Soc. Am. **89**, 1364–1377.

Summerfield, Q. (**1992**). Personal communication.

Weintraub, M. (**1985**). "A theory and computational model of auditory monaural sound separation," Ph.D. thesis, Stanford Univ., Stanford, CA.

Zwicker, U. T. (**1984**). "Auditory recognition of diotic and dichotic vowel pairs," Speech Commun. **3**, 265–277.